LINE OF BEST FIT 6.1.1

The line of best fit can model an association between two variables. In this lesson, lines of best fit are estimated by "eyeballing" them on a scatterplot. Technology is used in lessons later in this chapter.

Once the line of best fit is established, a slope and y-intercept can be established. In statistical analyses, the slope is often written as the amount of change one could expect in the dependent variable (Δy), when the independent variable is changed by one unit ($\Delta x = 1$). The y-intercept is the predicted value of the dependent variable when the independent variable is zero. Be careful. In statistical scatterplots, the vertical axis is often not drawn at the origin, so the y-intercept can be someplace other than where the line of best fit crosses the vertical axis of the scatterplot.

Associations between two variables are often described by their form, direction, strength of association, and outliers. Form refers to the shape of the pattern in the scatterplot: linear, some type of curve, or perhaps no pattern at all. Direction refers to whether the pattern is in general increasing (from left to right) or decreasing. For linear relationships, the slope can be used to describe the steepness in addition to the direction. The strength of the relationship indicates how closely the data conforms to the established form. Outliers are data points far removed from the bulk of the data.

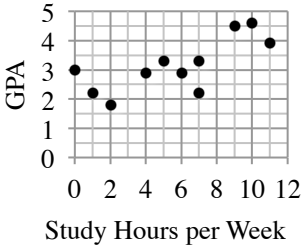
For additional information, see the Math Notes boxes in Lessons 4.1.1, 4.2.1, and 6.1.2, and the narrative in Checkpoint 8 at the back of the textbook. (Note that the problems below do not use a graphing calculator, while the problems in Checkpoint 8 assume a graphing calculator is available. Graphing calculators are introduced for statistical calculations in Lesson 6.1.4.)

Example

It seems reasonable that there would be a relationship between the amount of time a student spends studying and their GPA. Suppose you were interested in predicting a student's GPA based on the hours they study per week. You were able to randomly select 12 students and obtain this information from each student.

Hours	4	5	11	1	15	2	10	6	7	0	7	9
GPA	2.9	3.3	3.9	2.2	4.1	1.8	4.6	2.9	2.2	3	3.3	4.5

a. Without a calculator, make a scatterplot.



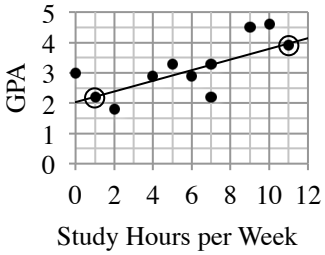
Example continues on next page →

© 2013 CPM Educational Program. All rights reserved.

Example continued from previous page.

b. Estimate a line of best fit (use a ruler and "eyeball" it), and determine its equation.

If it is reasonable, draw the line of best fit through a lattice point (where grid lines intersect) or a data point. That way, the equation will be easier to write. In this problem, the data points (1, 2.2) and (11, 3.9) seem to be on a line that "fits" the data.



Using those two points, and techniques from Lesson 2.3.2, the equation for the estimated line of best fit is $y = 2.03 + 0.17x$, where y is the predicted GPA and x is the number of hours per week the student studies. Note that the variables and their units were defined.

c. Interpret the slope and y-intercept in terms of study hours and GPA.

The slope indicates that a student's GPA is expected to increase by 0.17 points for every additional hour of studying per week. The y-intercept predicts that students who do no studying at all have a GPA of 2.03.

d. Describe the association.

To describe an association means to describe the form, direction, strength, and outliers. The form looks to be linear; it does not appear to be curved nor simply a collection of randomly scattered points. The direction is positive; as we find students who study more, we also find higher GPAs. A student's GPA is expected to increase by 0.17 points for every additional hour of studying per week. The strength is moderate: it is strong enough to easily see its form, but there is scatter about the line. There do not appear to be any outliers.

In summary one could say there is a moderately strong linear relationship between study hours and GPA for students with no outliers in our data.

Problems

1. It seems reasonable that the horsepower of a car is related to its gas mileage. Suppose a random sample of 10 car models is selected and the engine horsepower and city gas mileage is recorded for each one.

HP	197	170	166	230	381	438	170	326	451	290
MPG	16	24	19	15	13	20	21	11	10	15

a. Create a scatterplot of the data.

b. Estimate a line of best fit and determine its equation.

c. Interpret the slope and y-intercept in the context of horsepower and MPG.

d. Describe the association.

© 2013 CPM Educational Program. All rights reserved.

2. Many people believe that students who are strong in music are also strong in mathematics. But the principal at University High School wonders if that same connection exists between music students and English students. The principal went through the records for the past year and found 10 students who were enrolled in both Advanced Placement Music and Advanced Placement English. He compared their final exam scores.

   a. Create a scatterplot of the data.

   b. Estimate a line of best fit and determine its equation.

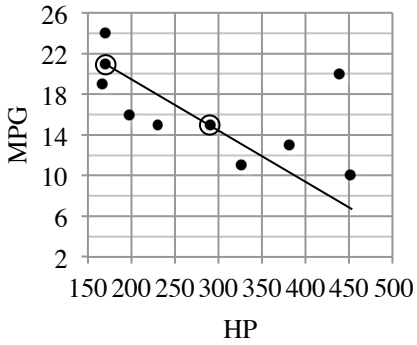
   c. Interpret the slope and y-intercept in context.

   d. Describe the association.

Final Exam Scores	
AP Music	AP English
88	63
74	96
82	86
64	90
97	68
90	90
82	78
72	74
78	96
62	79
checksum 789	*checksum 820*

Answers

1. a. See plot at right.



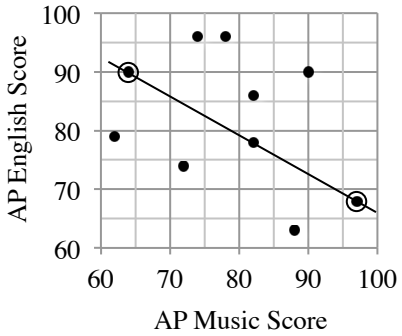
   b. Answers can vary. If you ignore the possible outlier at (438, 20), a reasonable line could pass through the points (170, 21) and (290, 15). From techniques in Lesson 2.3.2, the line of best fit has equation $y = 29.5 - 0.05x$.

   c. For every increase of one horsepower, gas mileage is expected to decrease by 0.05 mpg. The y-intercept means that a 0 horsepower car would get 29.5 mpg, which does not make sense. The y-axis is far outside of the data, so it represents an extrapolation. Prediction models of all types are unreliable when you extrapolate them.

   d. When the outlier at (438, 20) is removed, there appears to be a strong, negative, linear relationship. If the outlier is not removed, the association is more moderate (not as strong). For every increase of one horsepower, gas mileage is expected to decrease by about 0.05 mpg.

© 2013 CPM Educational Program. All rights reserved.

2. a. See plot at right.



b. Answers can vary. A reasonable line could pass through the points (64, 90) and (97, 68). From techniques in Lesson 2.3.2, the line of best fit has equation $y = 132.7 - 0.67x$.

c. An increase of 1 point for the music score results in a predicted decrease of 0.67 points for the English score. The y-intercept would mean that a student who scored 0 on the music test is expected to score 133 on the English test. This does not make sense in the given context. The y-axis is far outside of the data and so represents an extrapolation. Prediction models of all types are unreliable when you extrapolate them.

d. There is a lot of scatter around the line of best fit. The linear association is weak. An increase of 1 point on the music test results in a predicted decrease of 0.67 points on the English test. There are no apparent outliers.

© 2013 CPM Educational Program. All rights reserved.